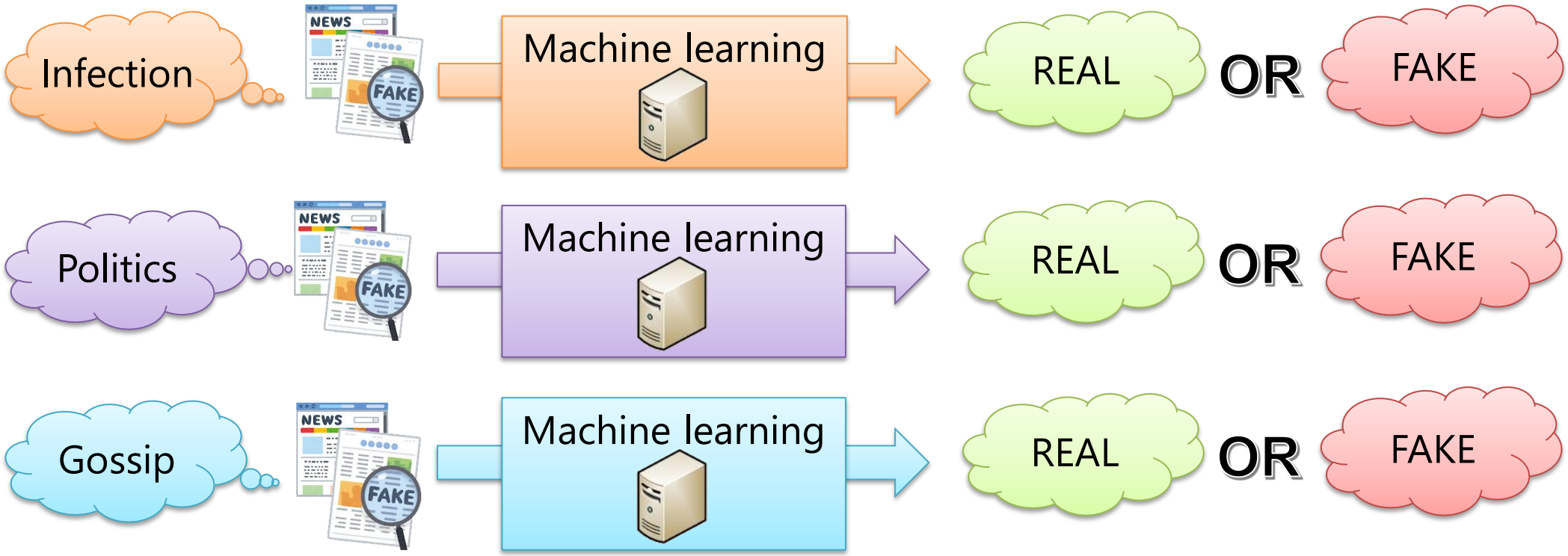
Extracting common features of fake news by Multi-Head-Attention

National Defense Academy of Japan
Takayuki Ishimaru
Mamoru Mimura

Contents

1． Background

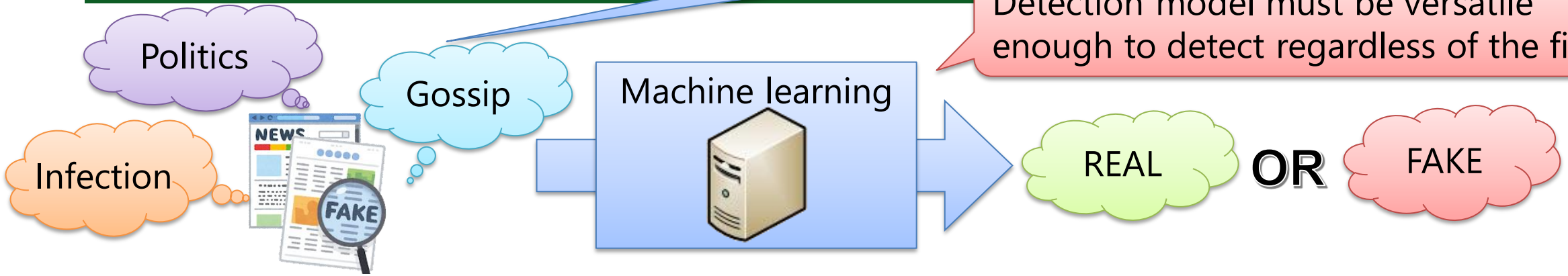
- Fake news has become a big problem for society
- Many methods have been proposed to detect fake news using machine learning models
- Most of those proposals have evaluated the classification accuracy for a specific dataset
- Few proposals for models that can deal with fake news in various fields with a single model.



1. Background

There is an unspecified large amount of information in social media

Detection model must be versatile enough to detect regardless of the field



Research objectives

- Extracting common features of fake news datasets
- Evaluating the versatility of the fake news detection model using BERT

Contribution

- Analysis of the words that Multi-Head-Attention focuses on
  ⇒ Confirmed that few words are common across datasets
- Evaluation of the fake news detection model by combining three different datasets, confirming that the model depends on the features of the training data
  ⇒ There is room for improving versatility

2. Related Work

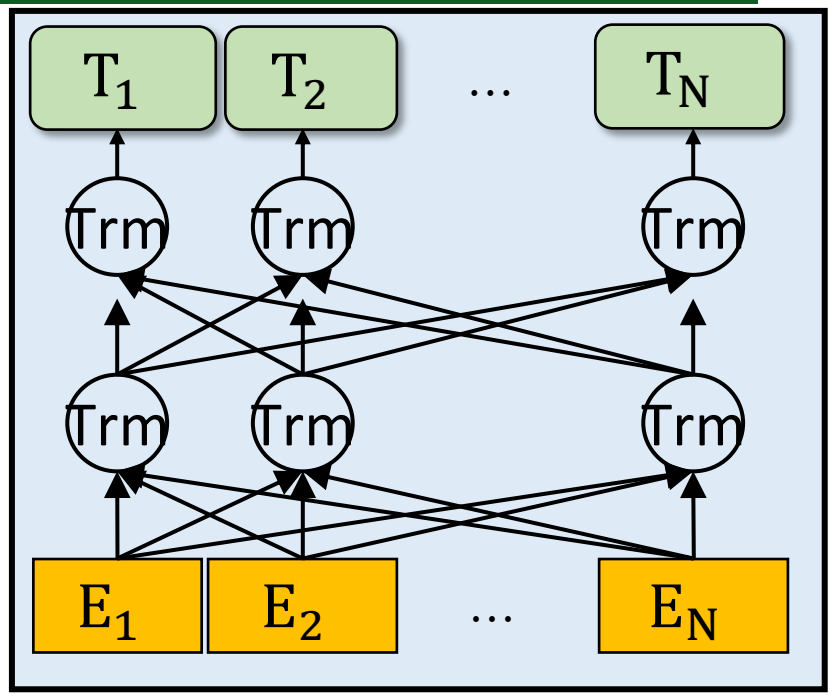
Many studies focus on word clouds or frequent words
⇒ Analysis the words to which the machine learning model directs its attention.

- Much research has been done on fake news detection
- Few detection models have been proposed to combine datasets and validate their generality

No.	Paper Title	Analysis of features	Dataset	Model
1	**Optimization and improvement of fake news detection using deep learning approaches for societal benefit** Tavishee Chauhan and Hemant Palivela International Journal of Information Management Data Insights(2021)	Word Cloud Word length	Kaggle.com	Glove-LSTM
2	**Deep contextualized text representation and learning for fake news detection** Mohammadreza Samadi, Maryam Mousavian, Saeedeh Momtazi Information Processing and Management(2021)	None	ISOT FAKENEWS Lair COVID-19	SLP, MLP, CNN in combination with BERT and GPT2
3	**Detecting English COVID-19 Fake News and Hindi Hostile Posts** Parth Patwa and Mohit Bhardwaj and Vineeth Guptha and Gitanjali Kumari and Shivam Sharma and Srinivas PYKL and Amitava Das and Asif Ekbal and Shad Akhtar and Tanmoy Chakraborty roceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation(2021)	Word Cloud Frequent vocabulary	COVID-19	SVM
	This Study	**Word analysis using BERT**	**ISOT FAKENEWS COVID-19 FA-KES**	**BERT**

3. Related Technique - BERT

- BERT (Bidirectional Encoder Representations from Transformers) is a natural language processing model.
- Designed to pre-train deep interactive representations from unlabelled text
- Highly pre-trained model and is known to have high accuracy in various tasks such as binary classification.
- Two BERT models, $\mathrm{BERT_{BASE}}$ and $\mathrm{BERT_{LARGE}}$, were proposed in Devlin et al



BERT pre-training model of Devlin et al.

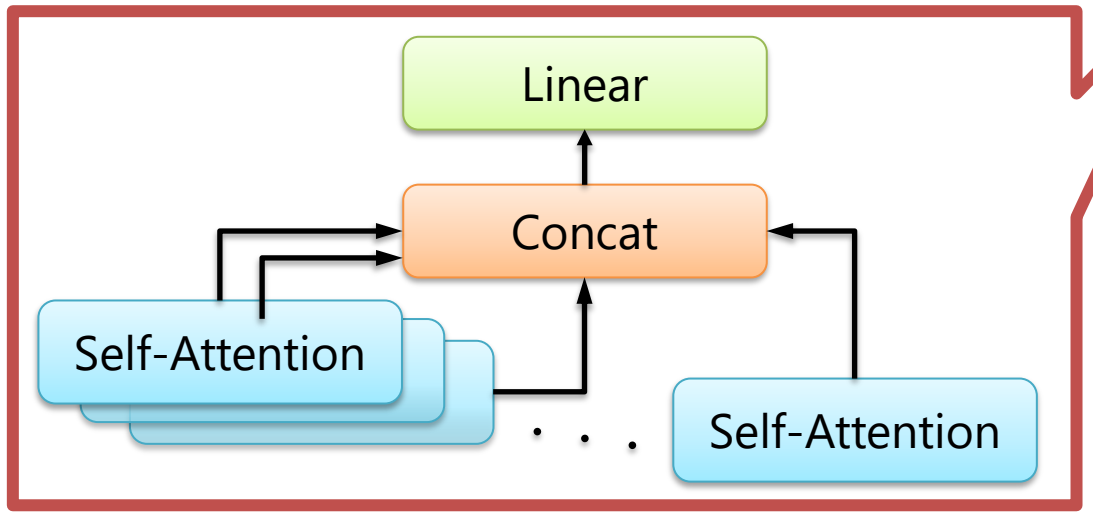
Parameter of BERT model

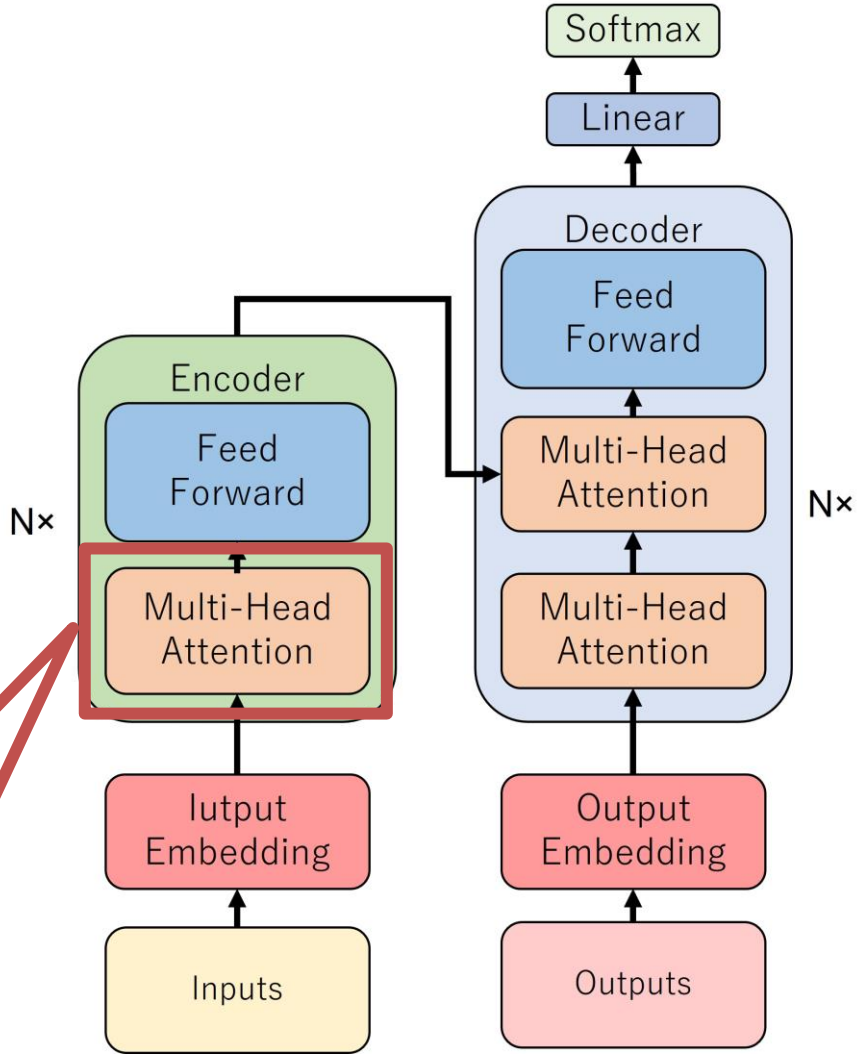
This study

Parameter name	$\mathrm{BERT_{BASE}}$	$\mathrm{BERT_{LARGE}}$
Layer num	12	24
Hidden layer	768	1024
Attention Heads	12	16
Parameter num	110M	340M

3. Related Technique - Multi-Head-Attention

- Part of a deep learning model called Transformer
- Acquire multiple Attention representations by computing multiple Attention in parallel
- Combined and transformed by the Linear layer to obtain the final attention representation
- We output Multi-Head-Attention weights to find out which words the machine learning model focuses on





Transformer model of Vaswani et al

3. Related Technique – Fine-tuning

・BERT has two stages: Pre-Training and Fine-Tuning

Pre-Training

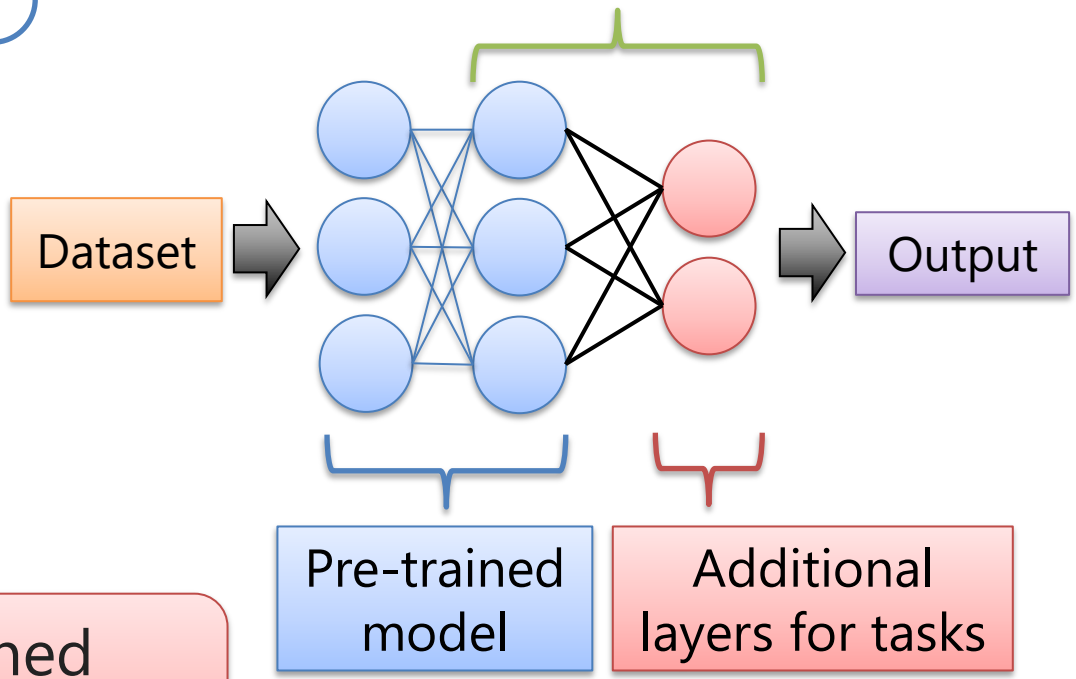
・Using MLM (Masked Language Model) and NSP (Next Sentence Prediction) to train on large amounts of unlabelled data
・Several pre-trained models have been published for each task

Fine-Tuning

・Add layers for a task to a pre-trained model
・Fine-tune the parameters for a specific task by training it with labelled data of the type you want it to learn

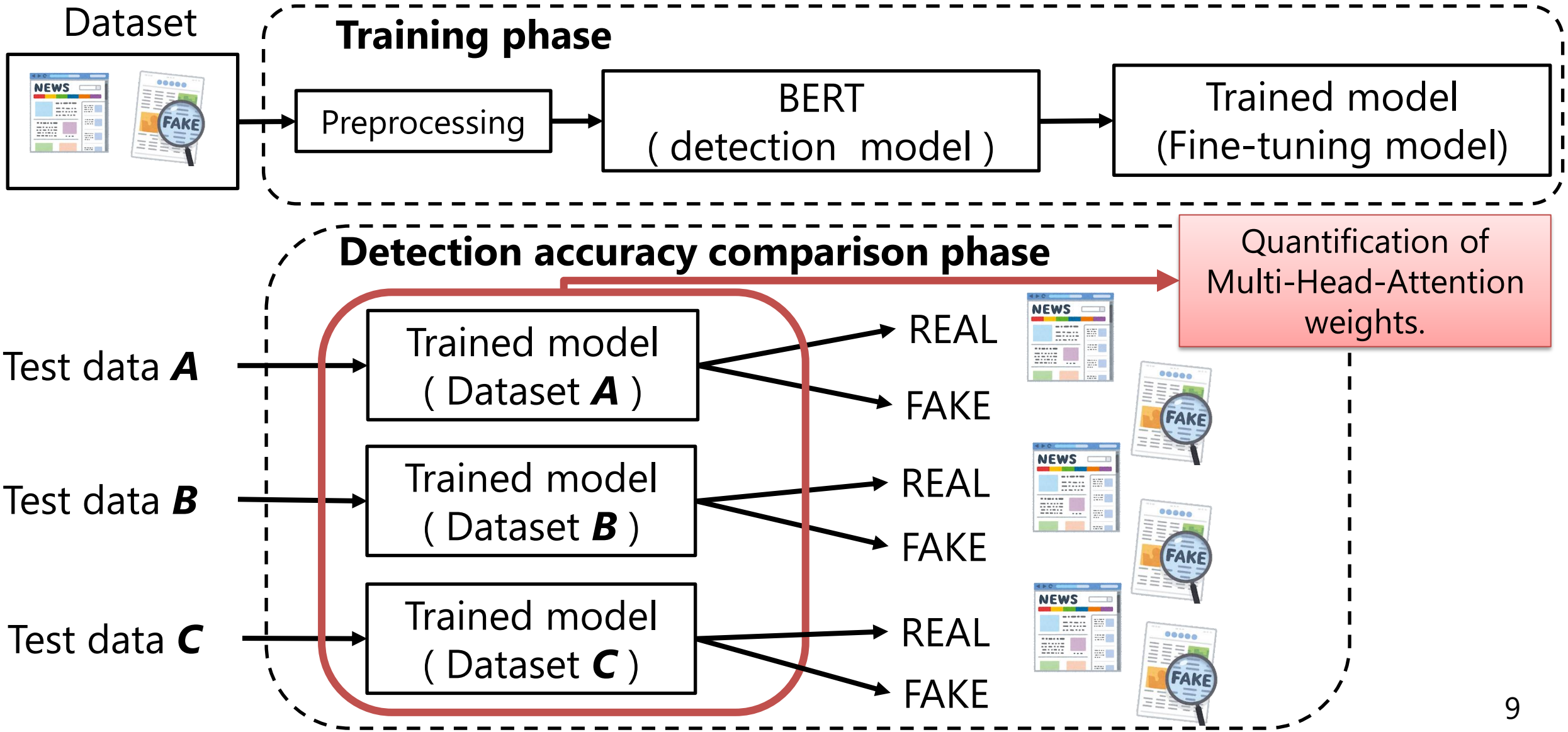
Highly accurate classification models can be obtained even with small amounts of training data.

Fine-tuning(updated), including parameters of pre-trained models

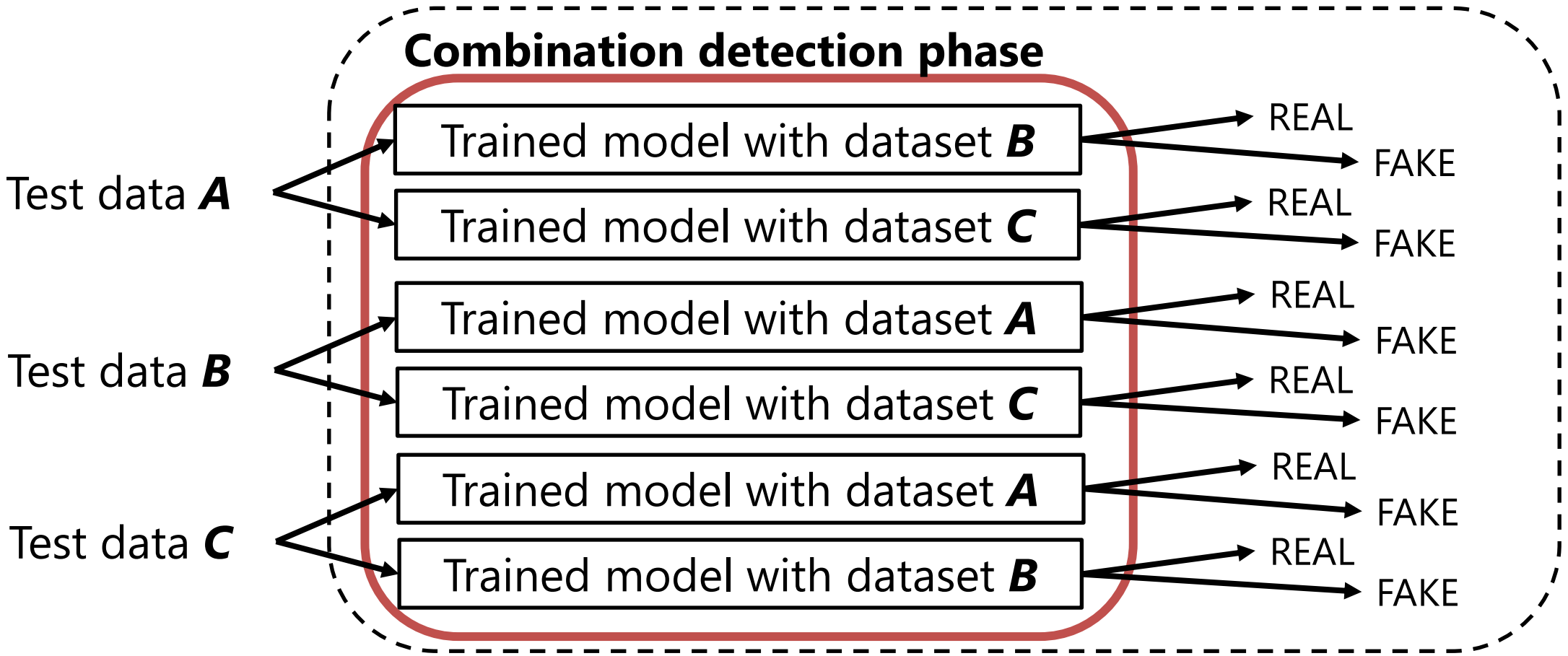


Examples of fine-tuning models

4． Evaluation method（1/3）



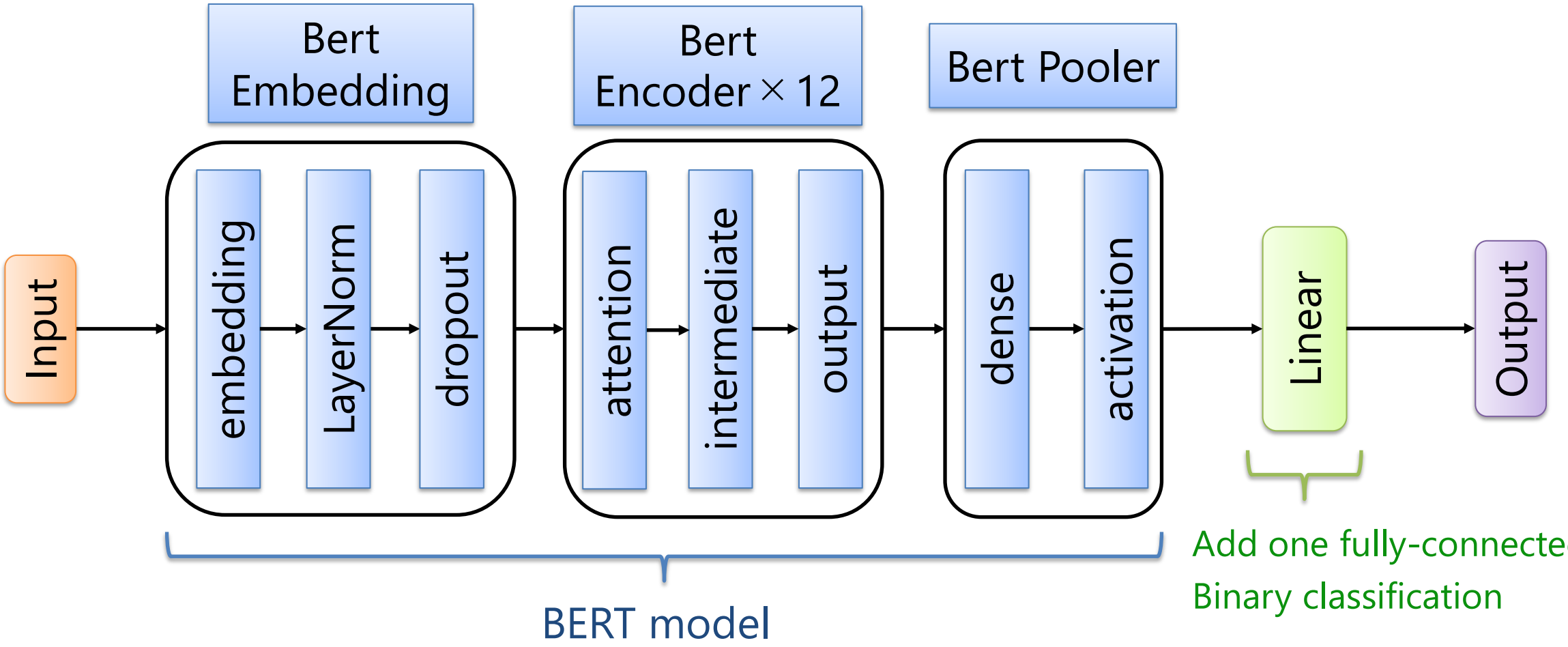
4． Evaluation method （2/3）



Check the detection accuracy in models trained on different types of datasets

4． Evaluation method（3/3）

• For evaluation, a BERT model with one additional fully-connected layer (Linear) for classification is used



5． Evaluation experiment - Dataset

- This study uses three datasets of different genres

Dataset name	Summary	FAKE	REAL
ISOT FAKENEWS	・Dataset on fake news in English ・Mainly global and political news	23481	21415
COVID-19	・Dataset on new coronavirus infections ・Posts and articles on Facebook, Twitter and other social media	5100	5600
FA-KES	・Dataset on the Syrian conflict	378	426

5. Evaluation experiment - Summary

• Summary of the experiment

No.	Summary	Dataset	
		Training	Test
1	・Check the accuracy of detection for each dataset ・For accuracy comparison with previous studies	ISOT	ISOT
		COVID-19	COVID-19
		FA-KES	FA-KES
2	・Extracting Multi-head-attention weights from each test data in the No.1 experiment ・The words in each text are sorted in order of weighted words, and the top five of these are extracted and aggregated	ISOT	ISOT
		COVID-19	COVID-19
		FA-KES	FA-KES
3	・Evaluation of the generality of the fine-tuning model ・Check the detection accuracy in models trained on different types of datasets	ISOT	COVID-19, FA-KES
		COVID-19	ISOT, FA-KES
		FA-KES	ISOT, COVID-19

5. Evaluation experiment - Environment

Environment

CPU	Core i7-9700K 3.60GHz
Memory	64GB
GPU	NVIDIA GeForce RTX 2080 Ti
OS	Windows10 Home
Programing language	Python3.8.9

Library

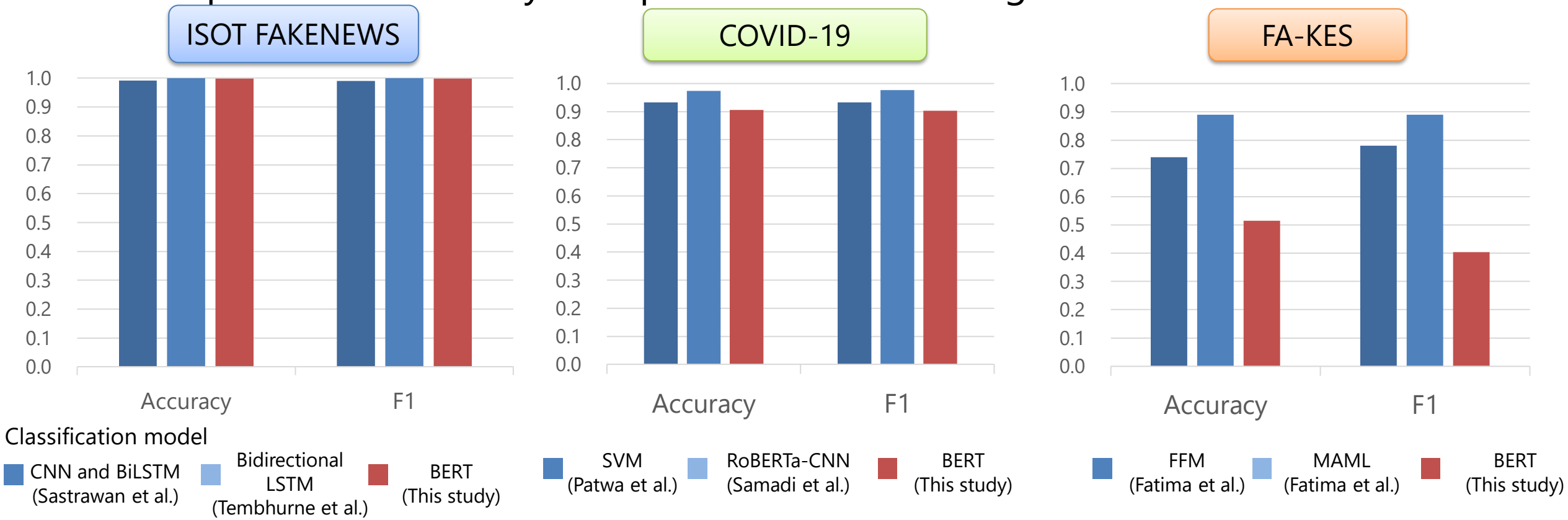
Machine learning library	PyTorch 1.7.1
	scikit-learn 1.0.2
Tokenizer	Wordpiece Tokenizer
Pretrain model	bert-base-uncaced
Vocabulary	bert-base-uncased-vocab

Hyperparameters

Dataset	ISOT	COVID-19	FA-KES
BATCH SIZE	128	64	16
EPOCHS	2	14	8
SEQUENCE LENGTH	256	50	256
Optimizer	Adam		
Learning Rate	0.00005		

5．Evaluation experiment – Result (No.1)

- Comparison of accuracy with previous studies using the same dataset



- ISOT FAKENEWS classified as comparable to previous studies
- COVID-19 and FA-KES were less accurate in this study
- FA-KES is hardly classifiable

5. Evaluation experiment – Result (No.2)

- The table shows the top 40 words extracted from the text identified as TP, out of the words that were the focus of attention in Multi-Head-Attention by dataset （**Blue word** indicates overlap between datasets）

No.	ISOT	COVID-19	FA-KES
1	**trump**	corona	**##s**
2	**said**	##vid	al
3	**would**	##virus	reported
4	**##s**	19	**##e**
5	hillary	**trump**	2011
6	com	**##e**	damage
7	time	**people**	conflict
8	**##e**	##19	**said**
9	yo	lock	killing
10	**##t**	**says**	group

No.	ISOT	COVID-19	FA-KES
11	sh	##de	**would**
12	**one**	president	##a
13	**says**	co	observatory
14	via	shows	government
15	like	claim	10
16	**th**	##o	source
17	**people**	**##s**	**people**
18	new	**w**	control
19	##ed	pan	statement
20	featured	india	**also**

- The results show that the top words are those that describe the characteristics of each dataset

5 . Evaluation experiment – Result (No.2)

No.	ISOT	COVID-19	FA-KES
21	donald	##n	monday
22	watch	##p	##or
23	states	fact	countryside
24	**##ing**	facebook	injured
25	##st	**said**	casualties
26	back	**##t**	according
27	bu	vaccine	##ad
28	obama	china	**says**
29	**also**	b	**th**
30	**twitter**	test	nu

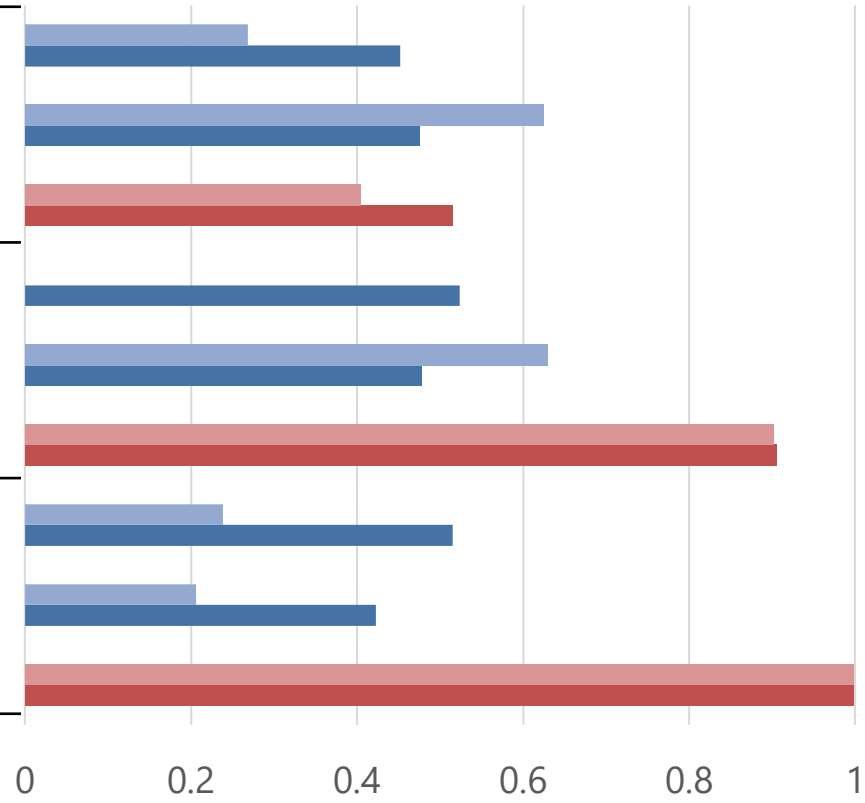
No.	ISOT	COVID-19	FA-KES
31	candidate	italy	##r
32	years	news	**twitter**
33	times	virus	attack
34	aft	health	**ou**
35	**w**	media	**##t**
36	**ou**	testing	houses
37	clinton	video	members
38	republican	##han	11
39	21st	patients	**one**
40	press	**##ing**	2016

Of the total 120 words, 15 words overlap between datasets, representing 14 % of the total

5. Evaluation experiment – Result (No.3)

- The results of the evaluation of the combination of Test data and Trained model

Test dataset	Trained model
FA-KES	COVID-19
	ISOT
	FA-KES
COVID-19	FA-KES
	ISOT
	COVID-19
ISOT	FA-KES
	COVID-19
	ISOT



・ F1 and Accuracy in Combination detection (result of No.3)

F1

Accuracy

・ F1 and Accuracy in result of No.2

F1

Accuracy

- No combination improved both F1 scores and Accuracy

6. Discussion (1/2)

Versatility of BERT fine-tuning model

- In BERT's fine-tuning model, there are high and low accuracy datasets
  ⇒ Limited types of detectable datasets

- As a result of combining the trained model and test data and evaluated, almost all combinations decreased detection accuracy

There is room for improvement regarding the versatility of the BERT fine-tuning model

6. Discussion (2/2)

Commonalities of Fake News

- Comparing the top 40 words in each dataset, 120 words in all, there are 15 words that overlap between datasets, with a proportion of 14 %

- Words in line with the characteristics of each dataset tend to rank higher

There are few common features of datasets that can be extracted using Multi-head-attention

7． Conclusion

- Focusing on the Multi-Head-Attention weights of the BERT model, an approach to extracting common features of the fake news dataset was attempted.

- Different datasets have different words that Multi-head-attention focuses on, indicating that there are few common features in fake news that can be extracted using Multi-head-attention

- By combining three datasets with different features, we show that fake news detection using fine-tuning model of BERT depends on the features of the training data

Thank you.